Measurement

Michael Coppedge
Varieties of Democracy Data Release Workshop
University of Notre Dame
January 22, 2016
Measurement model team

• Daniel Pemstein, Project Manager
• Kyle L. Marquardt, Research Fellow
• Eitan Tzelgov, former Research Fellow
• Yi-ting Wang, former Research Fellow
• Farhad Miri, Data Manager

Typical expert-rating projects

• Assume that experts rate without error
  • All interpret ordinal thresholds the same way: your “2” is the same as my “2”
    • even if they are coding different countries.
  • All experts are either
    • Perfectly skillful (when there is one expert)
    • or equally skillful (when there are multiple experts)

• V-Dem knows these are not safe assumptions.
Measurement challenges

• Some coders are less reliable than others.
  • Differences in amount of knowledge
  • Differences in type of knowledge
  • Differences in diligence: time spent, care, precision

• Differential Item Functioning (DIF)
  • Which information is relevant for answering this question?
  • How should I interpret the thresholds between the ordinal scores?

• Coders of the same country interpret our ordinal scales differently.
• Coders of different countries may interpret the scales differently.
We assume that coders/raters perceive a continuous underlying reality.
However, raters who perceive the same reality. . .
but with different ordinal thresholds...
can express their perceptions differently.
The result:
• It’s also possible that raters who *agree* on their observed ratings perceive different realities!

• So it’s very important to get good estimates of raters’ thresholds on each indicator.

• Dan Pemstein custom-designed a Bayesian Ordinal IRT measurement model to estimate these and other parameters.
Latent variables

- These are a class of models in which only some variables are observed (or “manifest”); others are unobserved (or “latent”).
- E.g., factor analysis, principal components
- Typically,

These can be estimated as a set of simultaneous equations: one for $X_1$, one for $X_2$, one for $X_3$. 
The Item-Response Theory (IRT) Framework: A special type of latent-variable model

We can understand observed ordinal scores as being above or below a threshold on a latent variable.

The higher the threshold, the more democratic the perceived reality must be to earn a higher ordinal score from the coder.
Ordinal IRT: 2 thresholds dividing 3 levels

Here there are ranges of the latent variable that correspond to each ordinal score:
0: less than -1
1: -1 to 0.6
2: greater than 0.6
Two parameters in ordinal IRT models

• **Difficulty** is estimated by the thresholds on the latent variable that separate ordinal scores. Each level of an indicator (minus one) has its own threshold.

• **Discrimination** is a coefficient estimating how crisply the coder distinguishes between ordinal scores. It determines the slope of the S-curve.

[See IRT simulator]
Bayesian estimation

• In a country-year-coder*indicator dataset, most of the cells would be empty because experts code only a few surveys in one or a few countries.

• Bayesian estimation avoids making the heroic assumptions that would be necessary using frequentist methods with such a sparse dataset.

• It also – through the magic of resampling – gives us confidence bounds around our parameters, including the latent variable.
The model estimates difficulty thresholds, assuming

- Global mean thresholds are between -2 and 2 (uniformly distributed)
- The mean country thresholds are allowed to vary around the global thresholds, with a standard deviation of 0.2
- Coder thresholds are allowed to vary around their country’s thresholds, with a standard deviation of 0.2

What this looks like:
An example for v2svinlaut: International autonomy.

Black: posteriors of global mean thresholds
An example for v2svinlaut: International autonomy.

Gold: 20 posteriors for all country thresholds
An example for v2svinlaut: International autonomy.

Blue: posteriors of coder thresholds for Denmark

Red: posteriors of coder thresholds for Venezuela
Why these assumptions?

• It’s a departure from the usual MCMC practice of weak priors, but much better than the typical expert-coding assumption that DIF is not an issue.

• It allows the lateral and bridge coding to help calibrate the thresholds.

• It helps especially with the countries that are not yet bridged, or not sufficiently bridged.
  • Ideally all countries would be connected, directly or indirectly, by a network of experts who have coded more than one country. This would enable us to compare a “3” in Gambia to a “3” in any other country.
This is the bridging as of March 2015 for the Elections survey. The bridging we need is nearly complete. Only 7 countries are not completely bridged.
Another issue

• Without a further assumption, latent variable estimates would be biased toward zero in some cases
  • Unbridged countries
  • Countries with invariant scores, which tell us nothing about what their coders’ thresholds would be for other scores

• The result is that the Switzerlands of the world would be biased downward and the Saudi Arabias would be biased upward. The model just wouldn’t “know” that a high score is really high and a low score is really low, so it would hedge its bets.
Fixed by an assumption about the latent variable:

• When the model does not have enough information about coders’ thresholds for a country, the country gets an average of the coders’ scores.*

• When the model does have enough information (which is most of the time!), this average is adjusted for the threshold estimates, as described above.

*Actually, the confidence-weighted average of the scores for all coders of that country-year, normalized with respect to scores for all country-years.
Temporal granularity

• Scores are not serially independent. The model does not assume that they are, so estimates are allowed to jump or fall suddenly when the data call for it.

• However, this falsely inflates the sample size, which would make us overconfident of the point estimates.

• Therefore, for the MM our observations are not country-years or country-days, but “regimes”: country-periods in which no coder changed his/her score or confidence for that country.

• This yields more conservative estimates.
Estimation

- Markov-chain Monte Carlo methods using Stan
- Iterative procedures identify the parameter estimates that best fit the observed data.
- We use high-performance computing hosted by the Center for Research Computing at Notre Dame.
- Each variable is modeled separately. It takes 2 hours to several days for each variable; weeks to estimate all 156!
- More detailed information is in Working Paper No. 21.
Validity and Reliability

Fernando Bizzarro Neto
Validity and Reliability

• Data must be valid (accurate) and reliable (consistent)
Validity and Reliability

- Data must be valid (accurate) and reliable (consistent)

- Checking for validity
  - Face Value
Freedom of academic and cultural expression (v2clacfree)
Is there academic and cultural freedom of expression related to political issues?
Is there academic and cultural freedom of expression related to political issues?

Responses:

0: Not respected by public authorities. Censorship and intimidation are frequent. Academic activities and cultural expressions are severely restricted or controlled by the government.

1: Weakly respected by public authorities. Academic freedom and freedom of cultural expression are practiced occasionally, but direct criticism of the government is mostly met with repression.

2: Somewhat respected by public authorities. Academic freedom and freedom of cultural expression are practiced routinely, but strong criticism of the government is sometimes met with repression.

3: Mostly respected by public authorities. There are few limitations on academic freedom and freedom of cultural expression, and resulting sanctions tend to be infrequent and soft.

4: Fully respected by public authorities. There are no restrictions on academic freedom or cultural expression.
Liberal Component (v2x_liberal)
To what extent is the liberal principle of democracy achieved?

• “Negative” view of political power: judges the quality of democracy by the limits placed on government.
Liberal Component (v2x_liberal)
To what extent is the liberal principle of democracy achieved?

• “Negative” view of political power: judges the quality of democracy by the limits placed on government.

\[ \text{Liberal Component} = \frac{\text{Rule of Law} + \text{Jud. Checks} + \text{Leg. Checks}}{3} \]

• Aggregation: Averaging across equality before the law and individual liberties (v2xcl_rol), judicial constraints on the executive (v2x_jucon), and legislative constraints on the executive (v2xlg_legcon)
Validity and Reliability

• Data must be valid (accurate) and reliable (consistent)

• Checking for validity
  • Face Value

• Checking for reliability
  • Confidence Intervals
We are not as confident about this datapoint as we are about this one. How does one incorporate this uncertainty in analysis?
V-Dem Workshop 2.0

• The method of composition
• Regression with uncertainty estimates
• And much more
Reliability of V-Dem Indices

• Internal Consistency
Reliability of V-Dem Indices

- Internal Consistency

- Bernhard et al. The Core Civil Society Index, V-Dem Working Paper 13
  - Factor Analysis

<table>
<thead>
<tr>
<th>Variable</th>
<th>Loading</th>
<th>Uniqueness</th>
</tr>
</thead>
<tbody>
<tr>
<td>v2cseeeorgs</td>
<td>0.96</td>
<td>0.07</td>
</tr>
<tr>
<td>v2csreprss</td>
<td>0.88</td>
<td>0.21</td>
</tr>
<tr>
<td>v2csprtcpt</td>
<td>0.82</td>
<td>0.31</td>
</tr>
</tbody>
</table>
Reliability of V-Dem Indices

• Internal Consistency

• Bernhard et al. The Core Civil Society Index, V-Dem Working Paper 13
  
  • Factor Analysis

<table>
<thead>
<tr>
<th>Variable</th>
<th>Loading</th>
<th>Uniqueness</th>
</tr>
</thead>
<tbody>
<tr>
<td>v2cseeorgs</td>
<td>0.96</td>
<td>0.07</td>
</tr>
<tr>
<td>v2csreprss</td>
<td>0.88</td>
<td>0.21</td>
</tr>
<tr>
<td>v2csprtcpt</td>
<td>0.82</td>
<td>0.31</td>
</tr>
</tbody>
</table>

• Other strategies: Correlations, ICC, Cronbach’s Alpha, Scatterplots

• Other Examples: Sundström et al: Women’s Political Empowerment, Sigman and Lindberg: Egalitarian Democracy
Comparisons

• V-Dem Party Institutionalization Index
Comparisons

• V-Dem Party Institutionalization Index
• Institutionalization very hard to measure, scholars have relied on indicators like electoral volatility or average party age
Comparisons

• V-Dem Party Institutionalization Index
• Institutionalization very hard to measure, scholars have relied on indicators like electoral volatility or average party age
Comparisons
Comparisons

Party system institutionalization index

- Western Europe
- South America

Rating

Comparisons

Party system institutionalization index

Western Europe
South America
Southeast Asia
Comparisons
Comparisons
Comparisons

- V-Dem Party System Institutionalization vs. Existing indirect measures of PSI
Shots fired: V-Dem x Polity
Shots fired: V-Dem x Polity

- Polity
- V-Dem
- Elec. Comp.
Alternative forms of the data

This would be a good time to open the dataset.
Relative scale

• Variables with no suffix: v2svinlaut, etc.
• The mean of many draws from the output
• Accompanied by *_codelow and *_codehigh bounds of the 70% highest posterior density (HPD) interval
• Best for most analyses: continuous, interval-level estimates
Ordinalized version

• Has the suffix *_ord
• The most probable original ordinal scale score (0, 1, 2, etc.) corresponding to the continuous MM estimates
• Includes *_ord_codelow and *_ord_codehigh HPD bounds, which are also integers.
• Appropriate if you need discrete indicators, for example for hazard rate models
Linearized Ordinal-Scale Posterior Prediction

• Also called “original scale” on the website
• Has the suffix *_osp and includes upper and lower bounds
• Intended to be the MM estimates rescaled to the original scale, but with degrees of closeness
• Calculated as a weighted average of each original score, weighted by the probability of that score.
• In line graphs, makes it easier to match scores to coding criteria.
• *Do not use in analyses:* not equal intervals; not necessarily closest to the most likely score.
Means

• Has the suffix *_mean
• Simply the unweighted average of all the ordinal scores coders submitted for each country-year or country-date
• Kind of a weird thing to do, but before the most recent version of the measurement model, some preferred it
• The sample size used to be larger because it yielded values even when there were too few coders.
K-chotomy classifications

- Have suffixes *3C, *4C, or *5C
- The relative scale values divided into 3, 4, or 5 ordinal categories
- Requested by those who want all variables recoded into the same number of categories
- Not recommended for most purposes
Planned improvements

• More lateral coding

• Historical V-Dem (Teorell, Knutsen, Gerring, Skaaning, Ziblatt, Cornell)
  • Back to 1789 or 1800, wherever possible
  • One expert per country, chosen for historical expertise

• Vignettes (Zimmerman, Glynn, Pemstein, Gerring)
  • The best way to anchor coder thresholds
  • This is being done for the 2016 update (in progress).
  • It asks experts to rate a pair of hypothetical vignettes on several key variables in each survey they do.
  • It does not cover all past coders, but will help. Eventually we hope to have almost all past coders answer the vignettes.